JOURNAL OF
MULTIDISCIPLINARY
SCIENCES AND INNOVATIONS

VOLUME 04, ISSUE 10
MOPNTHLY JOURNALS

ISSN NUMBER: 2751-4390
IMPACT FACTOR: 9,08

AI-BASED DETECTION AND CORRECTION OF LINGUISTIC ERRORS IN MACHINE TRANSLATION SYSTEMS THEORETICAL FOUNDATIONS

Shukurova Yulduz Yaxshimurotovna
Department of Foreign Language and Social Sciences,
Asia International University, English Teacher
Bukhara, Uzbekistan
E-mail: stellashukurova23@gmail.com

Abstract: This article examines the theoretical foundations of detecting and correcting linguistic errors in AI-based machine translation (MT) systems, with a focus on morphologically rich and low-resource languages such as Uzbek. Although neural architectures have advanced, MT outputs still contain grammatical, lexical, semantic, syntactic, and cultural errors. The paper analyzes key error sources and modern correction strategies, including rule-based methods, neural models, Quality Estimation (QE), and automatic post-editing (APE). Special emphasis is placed on phraseological units—idioms, proverbs, and fixed expressions—which remain challenging due to their figurative and culture-dependent meanings. The study also discusses the limitations of BLEU- and METEOR-based evaluation and highlights the importance of hybrid human–AI workflows. Future research directions include corpus expansion, improved semantic modeling, and linguistically informed neural techniques.

Keywords: Artificial intelligence, machine translation, linguistic errors, phraseological units, Uzbek, morphological complexity, deep learning, error correction, evaluation metrics, post-editing.

1. Introduction

Machine translation has rapidly evolved due to advances in artificial intelligence and deep learning. Modern systems such as Google Translate, Yandex Translate, DeepL, and ChatGPT generate increasingly fluent translations and are widely used across educational, industrial, and cross-cultural communication settings. However, the quality of such translations varies significantly depending on linguistic typology, availability of training data, and structural complexity of the language pair involved. Languages like English, German, and French-rich in digital resources-show relatively high translation accuracy, whereas low-resource, agglutinative languages like Uzbek present persistent challenges.

Uzbek, characterized by its extensive affixation, free word order in certain structures, and unique cultural expressions, often produces translation outputs with errors that range from minor grammatical issues to serious semantic distortions. These mistakes become more frequent when dealing with figurative language, idiomatic expressions, set phrases, and culturally embedded knowledge. Such linguistic units cannot be translated based on surface forms, and therefore require semantic reasoning, contextual understanding, and access to broader cultural references.

Given the growing reliance on machine translation in academics, journalism, medicine, law, and diplomacy, the need to detect and correct linguistic errors automatically is more urgent than ever. This article provides an extended theoretical discussion of MT error types, advanced AI-based detection mechanisms, and modern correction methods. It also highlights the special problems posed by phraseological units and outlines future prospects for MT in Uzbek and similar languages.

2. Types and sources of translation errors

Errors in MT output arise from system limitations, insufficient training data, typological mismatches, ambiguous structures, and contextual misunderstanding. Below are the major categories of errors expanded with detailed analysis.

2.1 Grammatical errors

Neural MT often fails to capture subject–verb agreement, tense consistency, case markers, and other morphological elements. Agglutinative languages like Uzbek express grammatical relations via long chains of suffixes (e.g., *ketayotganlaringizdanmisiz?*). These complex forms require morphological decomposition, which many systems struggle to perform.

Furthermore, definite and indefinite constructions, aspectual forms, and evidentiality markers in Uzbek frequently get mistranslated into English due to the lack of direct equivalents. For example, the Uzbek suffix *-ib* (indicating sequential actions) is often lost, causing semantic flattening.

2.2 Lexical errors

Lexical errors include using incorrect word choices, literal equivalents, or inappropriate synonyms. Polysemy is a major source: words like *yo'l* (road, path, method) or *chiqmoq* (exit, appear, publish) can be mistranslated if the context is not captured. Technical terminology also often receives incorrect interpretations because Uzbek lacks standardized translations in some scientific fields.

2.3 Semantic errors

Semantic errors occur when the system misinterprets relationships between concepts, misidentifies referents, or fails to maintain logical coherence. These include inconsistencies in number, time, possession, and modality. Semantic shift is especially common in idioms: for example, *ko'ngli qolmoq* is often translated as “his soul stayed,” rather than “to lose trust/interest.”

2.4 Syntactic errors

Uzbek's flexible word order allows topic-prominent structures and stylistic rearrangements. MT systems trained primarily on English-centric datasets often enforce English-like SVO patterns on Uzbek, distorting emphasis and meaning. Additionally, clause boundaries may be incorrectly aligned, especially when processing long sentences.

2.5 Stylistic and pragmatic errors

Stylistic errors involve tone, formality mismatches, unnatural phrasing, and loss of coherence. Pragmatic failure occurs when politeness levels, cultural nuance, or speaker intent is misinterpreted. Honorifics, modal nuances (*kerak, mumkin, shart*), and politeness markers often disappear in translation.

2.6 Cultural and idiomatic errors

Phraseological units represent one of the most problematic areas for MT. Idioms, proverbs, and metaphorical expressions encode cultural knowledge that neural networks may not access. For instance:

- *Tishi o'tmaydi* → “His teeth do not cut” (literal) instead of “He cannot cope with it”
- *Boshingga chiqmoq* → “To go on your head” (literal) instead of “To dominate or overpower”

These units require semantic interpretation rather than syntactic translation.

3. AI Methods for detecting linguistic errors

3.1 Rule-based approaches

Rule-based machine translation (RBMT) and grammatical checkers rely on handcrafted linguistic rules. Although limited, these systems excel in identifying morphological inconsistencies-an essential feature for Uzbek. They offer transparent explanations for detected errors and can complement neural outputs by providing structural constraints.

3.2 Statistical machine translation (SMT)

Although SMT has largely been replaced by neural approaches, its statistical alignment models help in diagnosing lexical and phrasing errors. SMT highlights common mistranslations by analyzing word frequency, alignment probabilities, and phrase tables, making it a useful error-analysis tool.

3.3 Neural architectures (NMT, Transformer, BERT, GPT)

Modern neural models outperform previous generations thanks to self-attention mechanisms and contextual embeddings. Transformers consider all words simultaneously, capturing long-distance dependencies. BERT-like models excel in classification tasks such as detecting errors, while GPT-based models excel in generating corrected versions.

However, these models struggle with:

- idiomatic expressions lacking training examples
- low-resource language embeddings
- morphological sparsity
- rare affix combinations
- cultural metaphors and pragmatic meaning

3.4 Quality estimation (QE) models

QE models evaluate translation quality **without reference translations**, predicting:

- sentence-level quality
- word-level errors
- gap insertions
- need for post-editing

They classify segments into “OK” or “BAD,” flagging problematic units for human editors. QE significantly reduces human workload.

3.5 Automatic post-editing (APE)

APE systems take raw MT output and produce a corrected version using a separate neural model. Studies show APE is particularly effective for repetitive errors or systematic morphological issues.

3.6 Reinforcement learning from human feedback (RLHF)

In RLHF, human editors rank or correct translations, and the model learns to prefer more accurate outputs. This method has significantly improved large language models used for translation.

4. Phraseological units in machine translation

Phraseological units (PUs) include idioms, proverbs, fixed expressions, collocations, and metaphorical constructions. Their translation requires deep cultural and semantic understanding-not merely lexical substitution.

4.1 Why MT struggles with PUs

- **Non-compositional meaning:** The meaning cannot be inferred from parts.
- **Cultural embeddedness:** Meanings depend on shared cultural knowledge.
- **Unpredictable structure:** Idioms may not follow typical grammar.
- **Cross-linguistic mismatch:** Many Uzbek idioms lack English equivalents.

4.2 Uzbek-specific PU challenges

Uzbek idioms frequently use body parts, symbolic categories, and metaphorical imagery:

- *Ko‘ngli tog‘dek bo‘ldi*
- *Yuragi orqasiga tushdi*
- *Qo‘lini sovuq suvga urmaydi*

Direct translation produces nonsensical output. Neural models are rarely exposed to idioms in sufficient quantity to generalize across contexts.

4.3 Improving PU translation

Solutions include:

- PU-rich corpora

- bilingual PU dictionaries
- semantic embeddings that detect figurative meaning
- idiom-aware pre-training tasks

5. Evaluation challenges and human-AI collaboration

5.1 Limitations of BLEU, METEOR and others

BLEU counts n-gram overlaps but fails to evaluate:

- paraphrasing
- idiomatic expressions
- culturally correct equivalents
- syntactic variability

A translation may be correct but receive a low BLEU score because it uses different wording. METEOR performs slightly better but still struggles with morphology-heavy languages.

5.2 Human evaluation

Human annotators assess:

- adequacy
- fluency
- cultural fit
- stylistic appropriateness

Human evaluation remains the gold standard.

5.3 Hybrid AI-human models

The best systems combine:

- neural MT
- rule-based morphology
- human post-editing
- QE
- RLHF

Such hybrid systems are especially effective for Uzbek.

6. Conclusion

AI-based machine translation continues to advance, but significant obstacles remain-particularly for morphologically complex, low-resource languages and idiom-heavy texts. Uzbek presents unique challenges due to its rich morphology, cultural nuance, and flexible syntax. Phraseological units remain particularly difficult for MT systems to handle, reinforcing the need for deeper semantic modeling and culturally informed datasets.

Future research must focus on:

- constructing large, diverse Uzbek corpora
- developing idiom-focused datasets
- integrating morphology-aware neural architectures
- advancing QE and APE models
- expanding human-in-the-loop systems
- improving evaluation metrics beyond surface-level comparison

References

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. NeurIPS.
2. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. ICLR.
3. Specia, L., Scarton, C., & Paetzold, G. (2018). Quality Estimation for Machine Translation. Morgan & Claypool. QE (xato aniqlash) bo‘yicha asosiy ilmiy manba.
4. Sennrich, R., Haddow, B., & Birch, A. (2016). Improving Neural Machine Translation Models with Subword Units. ACL.

5. Popović, M. (2020). Error Analysis in Machine Translation Output.

6. Freitag, M., Grangier, D., & Caswell, I. (2020). BLEU is Not Suitable for the Evaluation of MT for Morphologically Rich Languages.